CARDIOMNI: AN EVIDENCE-GROUNDED AGENTIC WORKFLOW FOR MULTI-VIEW MULTI-DIMENSION CARDIAC MRI INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cardiac magnetic resonance (CMR) interpretation spans a continuum from low-level visual perception to high-level clinical reasoning: a physician must inspect multi-view cine and LGE sequences, identify abnormal visual patterns, measure ventricular structure and function, and integrate these observations into a diagnosis. Existing CMR analysis systems are usually either task-specific predictors or end-to-end vision-language models that produce one-hop answers, making it difficult to accumulate explicit perceptual and quantitative evidence for downstream reasoning. Meanwhile, general and medical multimodal large language models have improved image-text understanding, but they are not designed to actively acquire CMR-specific evidence such as attribution maps, segmentation-derived volumes, ejection fraction, and indexed measurements. To bridge this gap, we propose **Cardiomni**, an evidence-grounded agentic workflow for multi-view, multi-dimensional CMR interpretation. Cardiomni coordinates two complementary evidence pathways through an iterative Planner–Executor–Verifier loop: a Visual Attention Highlighter provides abnormality probabilities with class-conditioned spatial evidence, while a Quantitative Measurement Estimator derives clinically interpretable measurements from cardiac masks and scan metadata. A persistent evidence ledger stores intermediate findings, enabling the planner to adaptively decide what evidence is still missing and allowing a frozen verifier and answer generator to produce answers grounded in the collected evidence. We train only the planner with reinforcement learning using a rule-based final-answer-correctness reward, while keeping the visual, quantitative, verification, and generation components fixed. We further introduce **DeepCMRVQA**, a benchmark for evaluating CMR perception-to-reasoning across 35 tasks covering major abnormality assessment and quantitative measurement interpretation. Experiments on binary abnormality VQA and measurement-range MCQ tasks show that Cardiomni outperforms general-purpose and medical vision-language baselines in balanced accuracy, highlighting a scalable path toward auditable CMR decision-support agents.

1 INTRODUCTION

Cardiac magnetic resonance (CMR) is a powerful imaging technique for assessing cardiac structure, function, tissue characteristics, perfusion, and blood flow, and is widely used in the diagnosis and management of cardiovascular disease Attallah & Ragab (2023). Unlike many medical imaging tasks that operate on a single image or scan (e.g., Chest X-ray Abnormalities detection), CMR is inherently multi-view and multi-sequence: a study may contain different anatomical views, cardiac phases, and sequences that provide complementary information, such as cine imaging for cardiac motion and ventricular function, late gadolinium enhancement (LGE) for myocardial scar, and perfusion imaging for blood supply. Interpreting CMR therefore requires selecting relevant images, performing quantitative measurements, and integrating evidence across sequences. However, the final radiology report typically summarizes the resulting findings and conclusions without fully preserving how the supporting images were selected, how measurements were obtained, or how individual findings were combined. This creates an important gap for machine learning: models trained pri-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

marily from images and reports can learn the final output, while the intermediate evidence-gathering and measurement process remains largely implicit.

Existing approaches fall short of this process in complementary ways. Task-specific deep learning systems can segment cardiac structures, calculate ventricular function, and classify disease Al-Antari et al. (2025); Ruijsink et al. (2020), and CMR foundation models such as ViTa Zhang et al. (2025b), CineMA Fu et al. (2026), and large-scale report-guided representation learning Shad et al. (2026) have extended this to a shared representation across multiple downstream tasks. However, their analysis steps are largely fixed in advance: they do not decide which measurement or attribution record a specific clinical question requires, and the connection between intermediate findings and the final conclusion often remains implicit. Recent agentic systems move toward more adaptive analysis. MESHAgents Zhang et al. (2025a) targets population-level phenotype association rather than patient-level CMR interpretation, while CARE Du et al. (2026) already includes dynamic planning, evidence–answer review, and reinforcement learning for selecting visual evidence. Cardiomni is therefore not distinguished by agentic planning alone; its focus is the CMR-specific choice between classifier attribution and reproducible quantitative measurement records derived from predicted cardiac masks. A major open challenge is to learn this selection process while keeping the final answer traceable to explicit evidence records rather than to an opaque forward pass. We discuss these and other related efforts in detail in Section 2.

We address this problem with **Cardiomni**, an evidence-grounded agentic framework for CMR interpretation. Given a clinical question, Cardiomni selects between the Visual Attention Highlighter and the Quantitative Measurement Estimator, stores the resulting observations in a structured evidence ledger, and asks a frozen verifier whether the ledger appears sufficient for answer generation. The ledger is designed for traceability: a generated answer can cite an existing classifier probability, attribution map, or quantitative measurement, and ventricular measurements can be recomputed from the predicted masks and scan metadata. This traceability is distinct from clinical correctness, which still depends on the quality of the masks, measurements, classifier outputs, and final interpretation. We formulate CMR interpretation as a sequential decision process in which the trainable planner proposes the next subgoal and evidence pathway from the clinical question, available pathway descriptions, and evidence collected so far. Cardiomni's classification and measurement backends remain fixed during agent training, while the planner is optimized with Group Relative Policy Optimization (GRPO) Shao et al. (2024) using a rule-based reward that scores only final-answer correctness. Routing quality, citation validity, and step efficiency are not direct reward terms in the present implementation, and we treat them as limitations and future extensions.

To support training and evaluation, we construct **DeepCMRVQA** for evaluating modern vision language model capability in multi-view multi-dimension cardiac MRI interpretation, across 35 tasks spanning cardiac structure, function, and imaging-feature assessment. These 35 evaluated tasks consist of 21 clinically central abnormality labels selected from the broader classifier label space for stronger positive-case support, together with 14 quantitative measurement metrics. Because our goal is to evaluate not only *what* the model predicts but also *which evidence it uses*, we report final-answer performance and analyze the behavior of the evidence-collection loop. Our main contributions are:

- We formulate comprehensive CMR interpretation as an adaptive evidence-collection process grounded in volumetric, multi-sequence quantitative evidence, rather than analyses fixed in advance or spatial localization alone.
- We introduce **Cardiomni**, which unifies classifier-derived abnormality evidence, segmentation-derived quantitative measurement, evidence construction, and answer generation within a single agentic workflow, with a structured evidence ledger that links each answer to explicit visual or quantitative evidence records.
- We combine supervised fine-tuned classification and measurement backends with an agentic decision policy trained via reinforcement learning against a verifiable, rule-based final-answer-correctness reward (§3.3).
- We present **DeepCMRVQA** across 35 evaluated tasks, comprising 21 major abnormality labels selected from a broader classifier label space and 14 quantitative measurement metrics; this paper reports binary cardiac abnormality classification and quantitative-measurement MCQ evaluation on a curated CMR mini test set.

2 Related Work

2.1 Agentic External Reasoning with Large Language Models

Large language models can augment parametric reasoning with external actions and tools, from ReAct, Toolformer, and Reflexion (Yao et al., 2023; Schick et al., 2023; Shinn et al., 2023) to large-scale tool-use benchmarks such as ToolLLM/ToolBench and τ-bench (Qin et al., 2024; Yao et al., 2024). More recent work directly optimizes agentic behavior with reinforcement learning: AgentFlow trains a planning policy within a modular agent loop (Li et al., 2026), while ARTIST learns when and how to invoke tools from outcome rewards (Singh et al., 2025). Cardiomni adapts this paradigm to cardiac MRI, where the planner selects between two clinically distinct and auditable evidence pathways—visual classification with attribution maps versus segmentation-derived quantitative measurements—and verifies whether the acquired evidence is sufficient before answering.

2.2 Medical Multimodal Large Language Models

Medical vision-language models such as LLaVA-Med, Med-Flamingo, MedGemma, and Lingshu improve multimodal clinical understanding but primarily reason from provided image-text context (Li et al., 2023; Moor et al., 2023; Sellergren et al., 2025; LASA Team et al., 2025). Medical-agent systems introduce external interaction: MMedAgent learns to select specialized medical tools (Li et al., 2024), while MedAgentBench and AgentClinic evaluate agents in interactive clinical environments (Jiang et al., 2025; Schmidgall et al., 2026). CARE further combines dynamic planning, explicit visual grounding, evidence review, and reinforcement learning (Du et al., 2026). In contrast, Cardiomni focuses specifically on cardiac MRI and learns to route questions between discriminative visual evidence and clinically interpretable measurements such as ventricular volumes, ejection fraction, and LV mass.

2.3 Reinforcement Learning with Verifiable Rewards

Reinforcement learning with verifiable rewards replaces learned reward models with automatically checkable outcomes. DeepSeekMath introduced Group Relative Policy Optimization (GRPO)(Shao et al., 2024), DeepSeek-R1 demonstrated large-scale RL-driven reasoning(DeepSeek-AI, 2025), and DAPO and Dr. GRPO further improve the stability and formulation of GRPO-style training (Yu et al., 2025; Liu et al., 2025). Recent work extends outcome-based RL to agentic trajectories involving tool use and planning (Singh et al., 2025; Li et al., 2026; Zhang et al., 2026). Cardiomni applies this setting to cardiac MRI, where final-answer correctness provides the reward for a Planner–Executor–Verifier trajectory over clinically meaningful evidence pathways, while the imaging tools and answer-generation components remain fixed.

3 Methodology

Given a medical question and visual inputs, Cardiomni does not run a fixed pipeline of analyses, nor does it stop at spatial localization. It treats interpretation as a process of **evidence collection**: at each step it decides what kind of evidence is still missing, routes the case through one of two dedicated evidence pathways to obtain it, writes the result into a shared **Evidence Ledger**, and asks a frozen **Evidence Sufficiency Verifier** whether the ledger already supports an answer. Once it does, a frozen **Summarizer** reads the ledger and produces the final diagnosis with explicit reference to the entries that justify it. We describe the two evidence pathways first (§3.1), then the ledger and the loop that fills and closes it (§3.2), then the reinforcement learning procedure that shapes the collection strategy itself (§3.3).

3.1 Evidence Pathways

Every finding that enters the Evidence Ledger originates from one of two evidence pathways. They are deliberately asymmetric: one provides classifier probabilities and attribution maps over a broad abnormality label set, while the other provides quantitative measurements deterministically computed from predicted cardiac masks and scan metadata. This asymmetry is the heart of the planner's job: each question must be matched to the form of evidence that can actually support it.

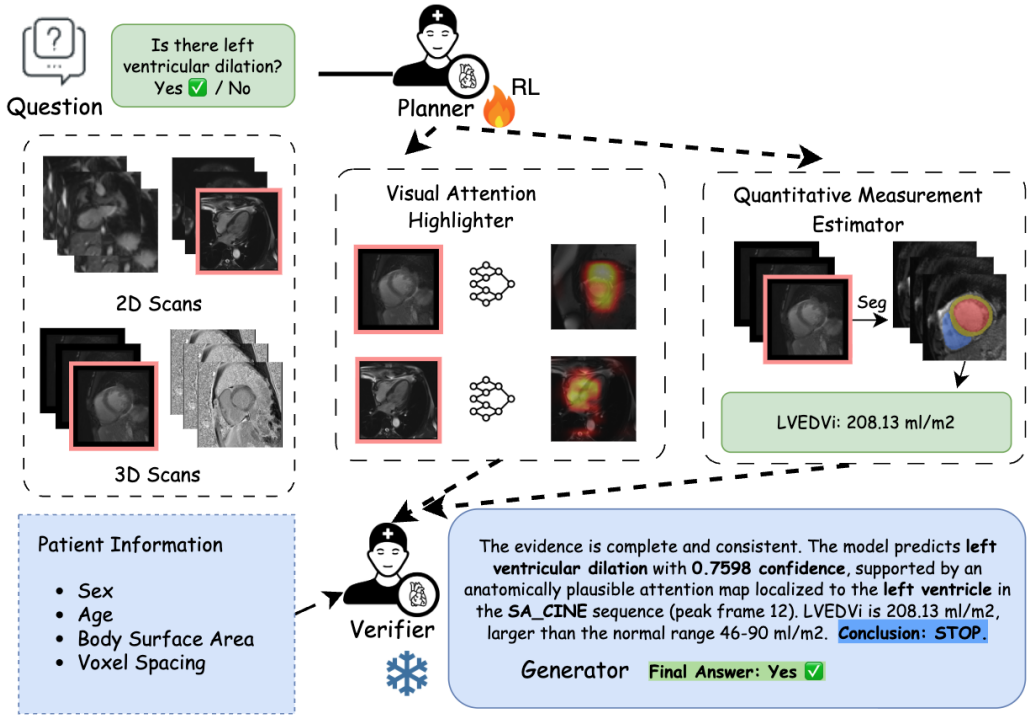


Figure 1: **Cardiomni workflow.** Given a clinical question, multi-sequence cardiac MRI, and patient metadata, Cardiomni iteratively builds an evidence ledger through an RL-trained planner. At each iteration, the planner decides which evidence is missing and routes the request to either the Visual Attention Highlighter, which returns abnormality probabilities with class-conditioned attention maps, or the Quantitative Measurement Estimator, which produces segmentation-derived cardiac measurements. A frozen verifier assesses whether the evidence ledger is complete and consistent; otherwise, the planner continues evidence collection. When the verifier judges the ledger sufficient, a frozen generator summarizes the ledger evidence into the final clinical answer.

Visual Attention Highlighter. This pathway localizes and characterizes abnormalities perceptually, without producing a physical measurement. A medical image classifier (Hara et al., 2018) $f_\psi : \mathcal{I} \to [0,1]^{32}$, with learned parameters ψ, maps a stacked cardiac-cycle clip $\mathcal{I}$ (of spatial resolution $H \times W$) to per-label probabilities over a label set $\mathcal{L}$, $|\mathcal{L}| = 32$. Its ledger-facing interface is summarized by a query operator $T_{\text{spat}}(\text{mode}, \ldots)$, whose first argument selects one of three modes and whose remaining arguments are mode-specific:

$$T_{\text{spat}}(\texttt{predict}, \ell) = f_\psi(\mathcal{I})_\ell, \quad \ell \in \mathcal{L}, \tag{1}$$

$$T_{\text{spat}}(\texttt{top_labels}, k) = \arg\operatorname*{top\text{-}k}_{\ell \in \mathcal{L}} f_\psi(\mathcal{I})_\ell, \tag{2}$$

$$T_{\text{spat}}(\texttt{attention_map}, \ell, \sigma) = \left(\text{CAM}^\star_{\ell,\sigma} \in [0,1]^{H \times W}, \ \{e_f\}_{f=1}^{F}, \ f^\star = \arg\max_f e_f \right), \tag{3}$$

where σ denotes a single sequence within $\mathcal{I}$ isolated for this query, and mode `attention_map` follows Grad-CAM (Selvaraju et al., 2017). The resulting overlay is class-conditioned evidence for an already requested label, not an independent detector: it can help the Verifier and Summarizer inspect whether a classifier probability is anatomically plausible, but it does not by itself establish that an abnormality is present. Appendix A.1 gives the full attribution derivation and downstream payload.

Quantitative Measurement Estimator. This pathway instead produces physically interpretable, independently re-derivable measurements from the short-axis cine stack. A semantic segmentation framework for biomedical images (Isensee et al., 2021) first delineates the left-ventricular blood pool, left-ventricular myocardium, and right-ventricular blood pool across cardiac phases, consistent

with Figure 1. The estimator then identifies end-diastolic and end-systolic phases from cavity-size extrema, converts segmented pixels into physical slice volumes using the scan spacing metadata, sums those slice volumes across the stack, and derives EDV, ESV, SV, EF, LV mass, and body-surface-area-indexed variants where applicable. The resulting ledger entries are numerical cardiac measurements, such as LVEF, LVEDV, LVESV, RVEDV, and LV mass, together with the reference ranges needed for quantitative MCQ reasoning. Full details of the phase selection, slice-wise volume computation, myocardial-mass conversion, and indexing procedure are given in Appendix B.2.

Why the pairing matters. Write $\mathcal{L}_{\text{spat}}$ and $\mathcal{L}_{\text{vol}}$ for the label sets addressable by the Visual Attention Highlighter and the Quantitative Measurement Estimator, respectively. The two evidence pathways are nested in coverage but inverted in evidentiary strength: $\mathcal{L}_{\text{vol}} \subset \mathcal{L}_{\text{spat}} = \mathcal{L}$, $|\mathcal{L}_{\text{vol}}| = 5$. For a label outside $\mathcal{L}_{\text{vol}}$, the Visual Attention Highlighter is the only route into the ledger; inside it, choosing between the fast perceptual pathway, the slower quantitative pathway, or both in sequence is a real decision with no universal fixed rule, and it is exactly this decision that the rest of the system is built and trained to improve.

3.2 Filling and Closing the Evidence Ledger

The Evidence Ledger $h_t = (a_1, s_1, c_1, o_1, \ldots, a_{t-1}, s_{t-1}, c_{t-1}, o_{t-1})$ is the single persistent record that both pathways write into and that every downstream judgment is read from: at step $\tau \in \{1, \ldots, t-1\}$, a_τ is the sub-goal pursued, s_τ the selected evidence pathway, c_τ the grounded request issued to that pathway, and o_τ the raw finding it returned. Nothing enters h_t except through this record, and nothing leaves the ledger once written, since the answer is required to reference it, never to bypass it.

At each step, the system decides what evidence to seek next and routes the case accordingly:

$$(a_t, s_t) = \text{select}(q, z_0, M(\mathcal{E}), h_t), \qquad c_t = \text{ground}(a_t, s_t), \qquad o_t = s_t(c_t), \tag{4}$$

where z_0 is a lightweight advisory parse of the question type and target label or metric produced from the question text alone, and $M(\mathcal{E})$ describes the two evidence pathways, including their input conventions, best-practice notes, and coverage limits. The planner uses these inputs to propose a natural-language subgoal and an evidence pathway; the grounding function then maps that proposal to concrete request arguments, such as an abnormality label for classification or a requested measurement metric. After each evidence request, the frozen Evidence Sufficiency Verifier inspects the ledger against the question and issues a binary judgment:

$$v_t = \text{sufficient}(q, h_{t+1}) \in \{\text{YES}, \text{NO}\}. \tag{5}$$

This judgment is an LLM-based sufficiency assessment, not a guarantee of clinical correctness. The loop runs until $v_t = \text{YES}$ or a step budget $T_{\max}$ is reached. We define the terminal step as $N = \min(\{t : v_t = \text{YES}\} \cup \{T_{\max}\})$, at which point the frozen Summarizer reads the available ledger and produces the answer with intended citations to ledger entries:

$$\hat{y} = \text{summarize}(q, h_{N+1}), \qquad \text{cite}(\hat{y}) \subseteq h_{N+1}. \tag{6}$$

This mechanism gives traceability, because an answer is required to reference entries already present in the ledger. It does not by itself prove that the cited entry supports the claim or that the underlying prediction is clinically correct. If v_t never reaches YES, the loop exits on the step budget alone and the Summarizer in Eq. equation 6 is invoked on whatever ledger exists at that point rather than the system declining to answer. There is currently no abstention option; every question receives a forced answer within $\{\text{YES.}, \text{NO.}\}$ or $\{A, B, C, D\}$, so a case that genuinely lacks sufficient evidence at $T_{\max}$ looks, at the interface level, identical to one the Verifier assessed as sufficient before the budget was exhausted. We return to this in §F.

Crucially, select is never given a hard-coded question-to-pathway table, and it must infer, purely from $M(\mathcal{E})$ and what the ledger already contains, whether the next best move is perceptual localization, quantitative measurement, or a second pass through a pathway already used because the first pass left the Verifier unsatisfied. The Sufficiency Verifier and the Summarizer stay frozen throughout; the only component subject to learning is this routing decision.

3.3 RL-BASED TRAINING WITH A VERIFIABLE REWARD

An untrained select can execute the entire loop above out of the box, but not well: it may choose the wrong evidence pathway or let the Summarizer cite an entry that was never really populated. We train select (planner parameters η) with reinforcement learning while ground, sufficient, and summarize remain frozen, using a rule-based final-answer-correctness reward with no learned reward model and no LLM-judge call.

Correctness. Let $y^\star$ be the ground-truth answer, drawn from a closed space ($\{\text{Yes.}, \text{No.}\}$ for VQA rows, $\{A, B, C, D\}$ for MCQ rows), and $\text{parse}(\hat{y})$ a deterministic extraction rule (last non-empty line, else a full-text unambiguous single-token scan). The reward is

$$R(\hat{y}, y^\star) = R_{\text{ans}}(\hat{y}, y^\star) = \mathbb{I}[\text{parse}(\hat{y}) = y^\star] \in \{0, 1\}, \tag{7}$$

following the verifiable-reward paradigm of DeepSeek-AI (2025): no learned reward model, no LLM-judge call, and no auxiliary term beyond final-answer correctness, since the current implementation does not score citation validity against h_{N+1}, reasoning consistency, or step count N directly. The parser extracts any valid answer token in the task's answer space independently of the ground truth; the equality check in Eq. equation 7 determines correctness after parsing.

Pathway selection is shaped implicitly, not scored directly. There is no ground truth for "the right pathway to follow," and R in Eq. equation 7 does not inspect h_{N+1} or $\{s_1, \ldots, s_N\}$ at all, since it depends on $\hat{y}$ alone. A wrong or wasted evidence request is therefore only ever penalized indirectly, through its downstream effect on whether $\hat{y}$ ends up correct (e.g., requesting `predict` where the question requires a quantitative EF value the Quantitative Measurement Estimator would have provided). This is a design limitation as much as a design choice: the reward cannot distinguish a correct answer reached via a well-justified routing sequence from one reached by a fortunate guess after an unnecessary or wrong evidence request, which is a natural direction for extending Eq. equation 7 with explicit terms for citation validity (R_{evi}, checking that $\text{cite}(\hat{y}) \subseteq h_{N+1}$ traces to populated entries) or step efficiency ($R_{\text{eff}}(N) = 1 - N/T_{\max}$) in future work.

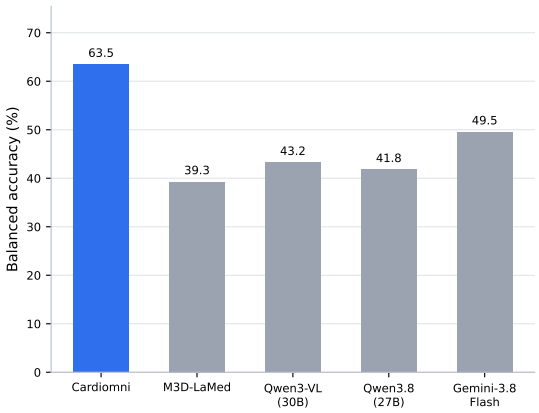


Figure 2: Average balanced accuracy across the DeepCMRVQA evaluation tables.

Optimization. We train the planner policy π_η with GRPO (Shao et al., 2024). For each question x, we sample G complete trajectories $\{\tau^{(1)}, \ldots, \tau^{(G)}\}$, where each trajectory contains the planner-generated subgoals/pathway choices, grounded evidence requests, verifier outcomes, ledger entries, and final answer. The frozen visual and quantitative evidence backbones, verifier, grounding routine, and Summarizer are excluded from the policy loss; only planner-generated tokens or action strings are scored. Each trajectory receives the final-answer reward in Eq. equation 7, giving $\{R^{(1)}, \ldots, R^{(G)}\}$, and the group-normalized advantage is

$$A^{(i)} = \frac{R^{(i)} - \text{mean}(\{R^{(j)}\}_{j=1}^{G})}{\text{std}(\{R^{(j)}\}_{j=1}^{G}) + \delta}, \tag{8}$$

with a small stabilizer δ; groups with zero reward variance therefore contribute zero centered advantage up to this stabilizer rather than producing an undefined update. Let $u^{(i)}$ denote the planner-generated token sequence or serialized action sequence in trajectory $\tau^{(i)}$, and let $c^{(i)}$ denote the corresponding planner context, including the question, pathway documentation, advisory parse, and ledger history visible at the decision points. The clipped planner objective is

$$\mathcal{J}(\eta) = \mathbb{E}_{x, \{\tau^{(i)}\}} \left[\frac{1}{G} \sum_{i=1}^{G} \min\Big(\rho^{(i)} A^{(i)}, \text{clip}(\rho^{(i)}, 1-\epsilon, 1+\epsilon)\, A^{(i)} \Big) - \beta\, \mathbb{D}_{\text{KL}}(\pi_\eta \,\|\, \pi_{\text{ref}}) \right], \tag{9}$$

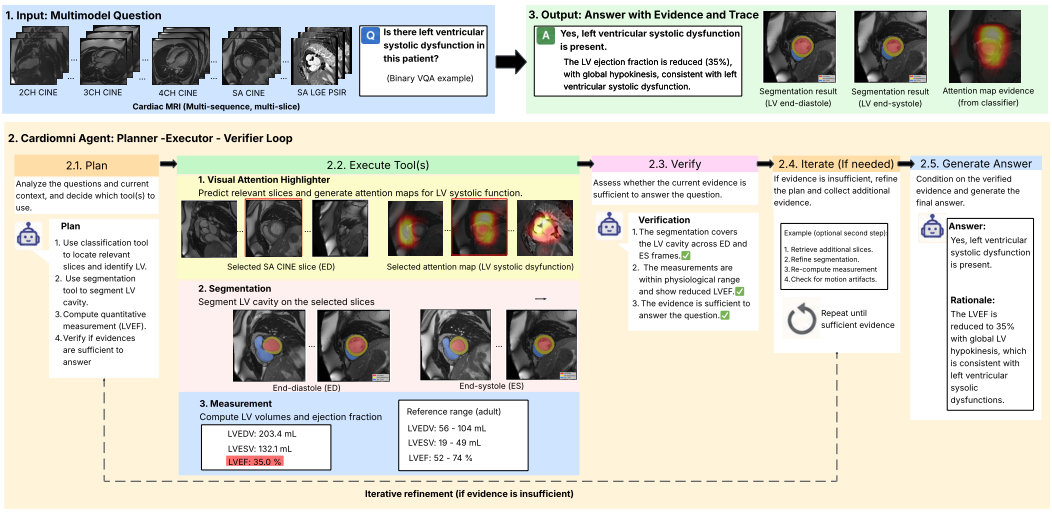


Figure 3: **Case study.** Given a cardiac MRI examination and a question about left ventricular (LV) systolic dysfunction, Cardiomni selects relevant images, highlights the LV, and segments it at end-diastole and end-systole. It calculates an LV ejection fraction (LVEF) of 35.0%, verifies the supporting measurements, including LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV), and concludes that LV systolic dysfunction is present. The final answer is supported by evidence extracted from the images and measurements.

where $\rho^{(i)} = \pi_\eta(u^{(i)} \mid c^{(i)})/\pi_{\eta^{\text{old}}}(u^{(i)} \mid c^{(i)})$. For an LLM planner, this ratio is implemented over planner-generated tokens with evidence records and frozen-backend responses masked from the loss. Because the reward is computed only from the final answer, observed gains may reflect changes in subgoal wording, pathway choice, or other planner outputs; we therefore avoid attributing improvements to pathway identity alone unless an ablation isolates that factor.

4 Experiments

4.1 Experimental Setup

Benchmarks. For evaluation, we use two task subsets from our constructed DeepCMRVQA benchmark. The binary VQA subset evaluates 21 major abnormality labels selected from the 32-label Visual Attention Highlighter output space; these labels were chosen to emphasize clinically central findings with stronger positive-case support. The quantitative MCQ subset covers 14 CMR-derived measurement metrics, where each question asks which of four sex-specific reference ranges a metric falls into. Unless otherwise specified, all experiments below use these same evaluation splits.

Comparative Methods. We benchmark Cardiomni against the medical-domain 3D vision-language model M3D-LaMed (Bai et al., 2024) and three general-purpose vision-language models, Qwen3-VL (30B) (Bai et al., 2025), Qwen3.8 (27B) (Qiu et al., 2026), and Gemini-3.8-Flash (Google DeepMind, 2026).

Fair Zero-Shot Prompting for Baselines. Because the zero-shot MLLM baselines have no access to Cardiomni's task-specific evidence pathways, evidence ledger, or training signal, we used a strong general-purpose LLM to author a shared evaluation harness. The harness specified question formats, answer conventions, and output parsing for both VQA and MCQ tasks, and was supplied verbatim to every baseline. This standardization reduces the risk that reported gaps reflect uneven prompt quality rather than differences in evidence-grounded reasoning.

Label	Cardiomni	M3D-LaMed	Qwen3-VL (30B)	Qwen3.8 (27B)	Gemini-3.8 Flash
aortic regurgitation	**67.3**	42.3	50.0	50.0	53.2
bicuspid aortic valve	**68.0**	–	49.7	50.0	46.3
cardiac tumor	**91.3**	50.0	49.0	50.0	49.0
delayed enhancement	59.3	48.1	**61.0**	54.0	50.5
left ventricular dilation	**79.3**	38.9	66.7	50.0	70.4
left ventricular systolic dysfunction	**68.7**	50.0	49.6	50.0	59.4
left ventricular wall motion abnormality	56.7	25.0	50.3	50.0	**59.8**
mitral regurgitation	**66.0**	50.0	50.0	50.0	51.5
myocardial edema	**74.0**	70.7	50.0	50.0	50.0
myocardial infarction	**83.3**	–	65.0	49.3	49.0
myocardial scar	**68.0**	49.1	61.2	45.7	57.2
no abnormal findings	**85.3**	50.0	49.7	50.0	44.1
pericardial abnormality	**74.7**	50.0	50.0	50.0	65.6
pericardial effusion	**77.3**	–	50.0	50.0	55.5
pericarditis	**90.7**	–	50.0	50.0	79.5
prior cardiac surgery	**83.3**	33.3	59.3	50.0	77.6
prosthetic heart valve	**69.3**	–	47.0	50.0	48.3
right ventricular dilation	**76.7**	49.4	54.2	49.6	59.3
right ventricular systolic dysfunction	**70.0**	50.0	50.0	50.0	57.5
right ventricular wall motion abnormality	58.7	50.0	50.0	50.0	**59.1**
tricuspid regurgitation	**66.0**	46.2	50.0	50.0	46.1
Average	**73.0**	47.1	53.0	49.9	56.6

Table 1: Binary VQA balanced accuracy by label. Best results are in **bold**; – denotes an undefined score.

Implementation Details. MLLMs are kept zero-shot in all experiments, with identical image inputs and, where applicable, matched prompt phrasing across both benchmarks. The general-purpose models are given up to 10 images per case, uniformly sampled at two frames per available sequence type. M3D-LaMed instead requires a single fixed-depth volumetric input, so each case's available sequences are concatenated along the depth axis into one $32 \times 256 \times 256$ volume, with frames budgeted proportionally across sequences to fill the fixed depth. Cardiomni's agent loop is capped at $T_{\max} = 5$ during training; reported evaluation permits up to eight evidence-collection steps, as detailed in Appendix C. We report balanced accuracy in both experiments below. Pathway quality, citation validity, and evidence sufficiency are not evaluated as separate reward-level metrics in the current experiments, consistent with the current reward's scope (§3.3, *Pathway selection is shaped implicitly*).

4.2 Comparison with MLLMs

We evaluate Cardiomni against M3D-LaMed and general-purpose vision-language baselines on two task families that map directly onto the two evidence pathways of §3.1. *Visual reasoning* probes the Visual Attention Highlighter's regime using binary questions across the 21-label evaluation subset. *Volume estimation* probes the Quantitative Measurement Estimator's regime using multiple-choice questions over 14 cardiac MRI–derived measurement metrics. Per-label and per-metric balanced accuracy are reported in Tables 1 and 2, respectively.

4.3 Analysis

Tables 1 and 2 show that Cardiomni is strongest on most visual abnormality labels and nearly all quantitative measurement metrics. The VQA table highlights the benefit of routing the question to label-specific classifier evidence and, when needed, attribution maps before answer generation: Cardiomni obtains the best balanced accuracy on 18 of the 21 evaluated findings, with especially large margins on tumor, pericarditis, prior cardiac surgery, no abnormal findings, and ventricular dilation. The few exceptions are findings where a direct VLM baseline is competitive, such as

Metric	Cardiomni	M3D-LaMed	Qwen3-VL (30B)	Qwen3.8 (27B)	Gemini-3.8 Flash
left ventricular end-diastolic volume	**50.4**	25.0	25.0	25.0	36.9
left ventricular end-diastolic volume index	**44.8**	25.0	25.8	26.7	29.5
left ventricular ejection fraction	**53.1**	41.6	25.0	29.4	31.5
left ventricular end-systolic volume	**52.7**	37.4	33.3	31.8	48.9
left ventricular end-systolic volume index	**50.9**	25.0	25.0	22.2	36.4
left ventricular mass	**49.3**	33.3	32.7	39.5	48.9
left ventricular stroke volume	**47.4**	25.0	25.6	23.2	28.2
left ventricular stroke volume index	**50.1**	22.8	25.0	22.2	30.7
right ventricular end-diastolic volume	**51.0**	33.3	33.3	33.3	49.2
right ventricular end-diastolic volume index	**50.3**	25.0	25.0	25.7	27.7
right ventricular ejection fraction	33.0	37.1	33.3	31.8	**42.0**
right ventricular end-systolic volume	**53.4**	32.1	33.3	42.1	53.2
right ventricular end-systolic volume index	46.2	33.3	33.3	31.0	**51.0**
right ventricular stroke volume index	**54.5**	29.2	25.0	29.5	30.9
Average	**49.1**	30.4	28.6	29.5	38.9

Table 2: MCQ balanced accuracy by measurement metric. Best results are in **bold**; random performance is 25.0.

delayed enhancement for Qwen3-VL and wall-motion abnormality for Gemini. In the MCQ table, the advantage shifts toward the quantitative pathway: Cardiomni is best on 12 of 14 metrics, while Gemini is strongest on RVEF and RVESVi and close on RVESV. This pattern is consistent with the intended division of labor in Cardiomni: direct VLMs can sometimes recognize salient visual patterns, but they are not designed to recompute ventricular volumes, indexed measures, and range assignments from physical measurements.

The many 50.0 balanced-accuracy values for Qwen3.8 (27B) in the VQA table mainly reflect conservative negative predictions rather than uniformly random behavior. It predicts NO for 4,604 of 4,800 questions (96%), close to the dataset skew of 4,413 NO labels (92%), but its YES recall is only 3.9%. Balanced accuracy is therefore more informative than plain accuracy for this imbalanced benchmark, because it exposes missed rare positive findings.

Why learned evidence routing matters. Cardiomni is designed around a routing problem rather than a fixed analytic recipe: the system must decide whether a question is best supported by visual attribution, quantitative measurement, or additional evidence after the ledger has already been partially filled. A fixed schedule can be effective when the evidence need is obvious and the available pathways are stable, but it does not express the broader goal of CMR interpretation: adapting evidence acquisition as the question, intermediate findings, and available evidence sources change. The learned `select` policy is therefore best viewed as an architectural mechanism for flexible evidence seeking. The same Planner–Executor–Verifier loop, evidence ledger, and GRPO training recipe can extend to settings with new imaging findings, additional measurement routines, or conflicting intermediate evidence, where the next useful evidence record cannot be specified by a static question-to-pathway table.

5 CONCLUSION

This work presents **Cardiomni**, an evidence-grounded agentic workflow for cardiac MRI question answering. Cardiomni routes each question to visual attribution or quantitative measurement evidence, stores the evidence in a persistent ledger, and trains the planner with a verifiable final-answer reward. Experiments on DeepCMRVQA show consistent gains over general-purpose and medical VLM baselines, especially on measurement questions that require explicit ventricular volumes and functional indices. These results suggest a practical direction for building CMR agents that are both accurate and traceable, while future work should further evaluate component contributions, citation quality, step efficiency, and abstention under insufficient evidence.

6 Ethics Statement

All authors have read and comply with the ICLR Code of Ethics. Private data used in this work were collected under institutional review board approval and de-identified prior to use. Cardiomni is intended only as a decision-support tool for qualified cardiac imaging clinicians, and its outputs should be reviewed before informing patient care. Potential risks include bias from uneven scanner, protocol, site, or patient-population coverage; cross-site generalization and subgroup fairness should be evaluated before clinical deployment.

7 Reproducibility Statement

We describe the Cardiomni architecture in Section 3, the planner training procedure and reward design in Section 3.3, and the evaluated DeepCMRVQA splits in Section 4. Additional implementation details for the visual and quantitative evidence pathways, language-model roles, preprocessing steps, and ablation setup are provided in the appendix. Where allowed by data-use agreements and privacy constraints, anonymized configuration files and processing scripts will be released as supplementary material.

AI Use Statement

In this work, we used generative AI tools to assist with drafting portions of the manuscript text and organizing the paper outline. We have not used generative AI tools to generate experimental results, and no results in this paper were produced or fabricated by a generative AI system. Where generative AI tools were used, all text and claims were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Mugahed A. Al-Antari, Riyadh M. Al-Tam, Aymen M. Al-Hejri, Zaid Al-Huda, Soojeong Lee, Özal Yıldırım, and Yeong Hyeon Gu. A hybrid segmentation and classification CAD framework for automated myocardial infarction prediction from MRI images. *Scientific Reports*, 15:14196, 2025. doi: 10.1038/s41598-025-98893-1.

Omneya Attallah and Dina A. Ragab. Auto-myin: Automatic diagnosis of myocardial infarction via multiple GLCMs, CNNs, and SVMs. *Biomedical Signal Processing and Control*, 80:104273, 2023. doi: 10.1016/j.bspc.2022.104273.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3d medical image analysis with multi-modal large language models, 2024. URL https://arxiv.org/abs/2404.00578. Introduces M3D-LaMed for 3D medical image analysis.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuexi Du, Jinglu Wang, Shujie Liu, Nicha C. Dvornek, and Yan Lu. CARE: Towards clinical accountability in multi-modal medical reasoning with an evidence-grounded agentic framework. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=whRAOJiyHM.

Delafield Du Bois and Eugene F. Du Bois. A formula to estimate the approximate surface area if height and weight be known. *Archives of Internal Medicine*, 17(6):863–871, 1916.

Yunguan Fu, Wenjia Bai, Weixi Yi, Charlotte Manisty, Anish N. Bhuva, Thomas A. Treibel, James C. Moon, Matthew J. Clarkson, Rhodri Huw Davies, and Yipeng Hu. Development and validation of a versatile foundation model for cine cardiac magnetic resonance image analysis. *Communications Medicine*, 6:408, 2026. doi: 10.1038/s43856-026-01636-0.

Google DeepMind. Gemini 3.8 Flash: Model card. Model card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018. doi: 10.1109/CVPR.2018.00685.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.510.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, 2023.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. In *International Conference on Learning Representations*, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: A multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, et al. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

Bram Ruijsink, Esther Puyol-Antón, Ilkay Oksuz, Matthew Sinclair, Wenjia Bai, Julia A. Schnabel, Reza Razavi, and Andrew P. King. Fully automated, quality-controlled cardiac analysis from CMR: Validation and large-scale application to characterize cardiac function. *JACC: Cardiovascular Imaging*, 13(3):684–695, 2020. doi: 10.1016/j.jcmg.2019.05.030.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9(1):499, 2026. doi: 10.1038/s41746-026-02674-7.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Rohan Shad, Cyril Zakka, Dhamanpreet Kaur, Mrudang Mathur, Robyn Fong, Joseph Cho, Ross Warren Filice, John Mongan, Kimberly Kallianos, Nishith Khandwala, David Eng, Matthew Leipzig, Walter R. Witschey, Alejandro de Feria, Victor A. Ferrari, Euan A. Ashley, Michael A. Acker, Curtis Langlotz, and William Hiesinger. A generalizable deep learning system for cardiac MRI. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01637-3.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool integration for LLMs via reinforcement learning. *arXiv preprint arXiv:2505.01441*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhong-Zhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Francisco Piedrahita Velez, Yue Liao, Hongru Wang, Mengyue Yang, Heng Ji, Jun Wang, Shuicheng Yan, Philip Torr, and Lei Bai. The landscape of agentic reinforcement learning for LLMs: A survey. *Transactions on Machine Learning Research*, 2026. URL `https://openreview.net/forum?id=RY19y2RI1O`.

Weitong Zhang, Mengyun Qiao, Chengqi Zang, Steven Niederer, Paul M. Matthews, Wenjia Bai, and Bernhard Kainz. Multi-agent reasoning for cardiovascular imaging phenotype analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15960 of *Lecture Notes in Computer Science*, pp. 429–439. Springer Nature Switzerland, 2025a. doi: 10.1007/978-3-032-04927-8_41.

Yundi Zhang, Paul Hager, Che Liu, Suprosanna Shit, Chen Chen, Daniel Rueckert, and Jiazhen Pan. Towards cardiac MRI foundation models: Comprehensive visual-tabular representations for whole-heart assessment and beyond. *Medical Image Analysis*, 106:103756, 2025b. doi: 10.1016/j.media.2025.103756.

APPENDIX

Appendix contents.

A	Additional Derivations
A.1	Full Grad-CAM Derivation for the Visual Attention Highlighter
B	Evidence Pathway Implementation
B.1	Visual Attention Highlighter
B.2	Quantitative Measurement Estimator
C	LLM Implementation
D	DeepCMRVQA Dataset Design
D.1	Image Preprocessing
E	Additional Results
E.1	Ablation Study
F	Limitation

A ADDITIONAL DERIVATIONS

A.1 FULL GRAD-CAM DERIVATION FOR THE VISUAL ATTENTION HIGHLIGHTER

This appendix gives the full derivation summarized by Eq. 3 in §3.1. For a query $T_{\text{spat}}(\texttt{attention_map}, \ell, \sigma)$, the classifier f_ψ is evaluated with every sequence input other than σ masked out, so that only σ's pathway contributes to the forward pass, and the logit for label ℓ is backpropagated. Let $A \in \mathbb{R}^{C\times F'\times H'\times W'}$ denote the forward activations at the classifier's final convolutional block (layer4), with C channels and reduced temporal/spatial dimensions F', H', W', and let $\partial \text{logit}_\ell / \partial A$ denote the backpropagated gradient at the same layer, of identical shape. Channel importance weights are obtained by global-average-pooling the gradient over the temporal and spatial axes,

$$w_c = \frac{1}{F'H'W'} \sum_{f=1}^{F'} \sum_{h=1}^{H'} \sum_{w=1}^{W'} \frac{\partial \,\text{logit}_\ell}{\partial A_{c,f,h,w}}, \qquad c = 1, \ldots, C, \tag{10}$$

and the raw class-activation map is their weighted sum over channels, followed by a rectification and a max-normalization to $[0, 1]$,

$$\text{CAM} = \text{ReLU}\left(\sum_{c=1}^{C} w_c\, A_c\right) \Big/ \left(\max \text{ReLU}\left(\sum_{c=1}^{C} w_c\, A_c\right) + \epsilon\right) \;\in\; [0,1]^{F'\times H'\times W'}, \tag{11}$$

with ϵ a small constant preventing division by zero. This raw map is then upsampled along both the temporal and spatial axes to match the input clip's native frame count F and resolution $H \times W$, giving $\text{CAM}_\uparrow \in [0, 1]^{F\times H\times W}$. Two derived quantities are read off $\text{CAM}_\uparrow$: a per-frame energy obtained by mean-pooling each frame's map over its spatial extent,

$$e_f = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (\text{CAM}_\uparrow)_{f,h,w}\,, \qquad f = 1, \ldots, F, \tag{12}$$

and the peak frame index $f^\star = \arg\max_f e_f$. The evidence actually deposited in the ledger is not the full tensor $\text{CAM}_\uparrow$ but a compact, inspectable triple: a single 2D overlay image $\text{CAM}^\star_{\ell,\sigma}$, formed by alpha-blending (0.55 source / 0.45 heatmap) the denormalized RGB frame $f^\star$ of $\mathcal{I}_\sigma$ with a colormap-rendered $(\text{CAM}_\uparrow)_{f^\star}$; the full per-frame energy sequence $\{e_f\}_{f=1}^{F}$; and the peak index $f^\star$ itself. Only $\text{CAM}^\star_{\ell,\sigma}$ is passed to the downstream Evidence Sufficiency Verifier and Summarizer as an image; $\{e_f\}$ and $f^\star$ are passed as text alongside it.

B EVIDENCE PATHWAY IMPLEMENTATION

B.1 VISUAL ATTENTION HIGHLIGHTER

The Visual Attention Highlighter is implemented with a 3D residual video classifier followed by class-conditioned Grad-CAM attribution. It predicts 32 report-derived abnormality labels from

multi-sequence CMR cine and LGE inputs. For each patient, up to five sequence groups are used: two-chamber cine, three-chamber cine, four-chamber cine, short-axis cine, and short-axis LGE/PSIR. Each available sequence is represented as a 16-frame RGB clip resized to 224×224, giving an input tensor of shape $B \times S \times 3 \times 16 \times 224 \times 224$, where $S \leq 5$ is the number of available sequence groups. All sequences are passed through a shared 3D backbone, unavailable sequences are masked out, and sequence-level features are aggregated with masked mean pooling before a dropout layer ($p = 0.3$) and a linear multi-label classification head.

The implementation supports two 3D residual backbones. The first is a 3D ResNet-152-style Bottleneck network with block depths $[3, 8, 36, 3]$, trained from scratch. The production classifier used for inference is based on the lighter R3D-18 architecture with BasicBlock depths $[2, 2, 2, 2]$ and Kinetics-400 initialization. The classifier was trained with AdamW, binary cross-entropy with logits, class-wise positive weights computed from the negative-to-positive ratio and clipped at 20.0, cosine learning-rate annealing, and bfloat16 mixed precision. Default optimization settings were learning rate 10^{-4}, weight decay 10^{-4}, batch size 4, a maximum of 30 epochs, and early stopping with patience 6.

Training data were constructed by merging two de-identified institutional label tables with the paired CMR image-report metadata. Labels with fewer than 20 positive cases were removed, yielding the final 32-label prediction space for the classifier. This 32-label space is broader than the 35-task DeepCMRVQA evaluation reported in the main text: the latter uses 21 clinically central abnormality labels with stronger positive-case support, plus 14 quantitative measurement metrics. The data were split using multilabel-stratified 8:1:1 train/validation/test partitioning with seed 42, resulting in 8,036 training, 1,004 validation, and 1,005 test patients. The production checkpoint was selected from an independent training run that stopped after 17 epochs, with the best validation macro average precision of 0.2770 reached at epoch 11. On the held-out test set, the classifier achieved macro AUC 0.8094 and macro average precision 0.2691, with per-label AUC ranging from approximately 0.50 for the most difficult labels to approximately 0.97 for systolic anterior motion of the mitral valve.

For attribution, we use the 3D Grad-CAM procedure described in Appendix A.1. Briefly, after a label is selected, gradients from the corresponding logit are backpropagated to the final convolutional feature tensor, channel weights are obtained by global-average-pooling the gradients, and a rectified weighted feature map is upsampled to the input frame resolution. The frame with the highest spatially averaged activation is used to produce the ledger-facing heatmap overlay. Thus, the Visual Attention Highlighter contributes both a calibrated abnormality probability and an anatomically inspectable explanation of the frames and regions that drove that probability, while remaining distinct from an independent localization or segmentation model.

B.2 Quantitative Measurement Estimator

The Quantitative Measurement Estimator is based on nnU-Net v2 2D segmentation followed by deterministic CMR volume measurement. Three independent segmentation datasets were prepared: ACDC 2017, M&Ms-2 short-axis, and M&Ms-2 long-axis. Cardiomni uses the M&Ms-2 short-axis model for quantitative measurement, because the downstream measurements are derived from short-axis ventricular stacks. Its label convention is background, left-ventricular cavity, left-ventricular myocardium, and right-ventricular cavity. The M&Ms-2 short-axis split contained 388 training/validation cases, partitioned into 310 training and 78 validation cases, with 314 held-out test cases. Nine short-axis patients with non-orthogonal NIfTI direction matrices were excluded during dataset preparation; the long-axis dataset also required conversion to a true 2D representation to avoid patch-size failures, although that model is not used for the reported short-axis measurements.

The segmentation backbone follows the nnU-Net v2 automatically configured 2D U-Net. The resulting network is a seven-level PlainConvUNet with feature widths $[32, 64, 128, 256, 512, 512, 512]$, 320×320 patches, batch size 31, instance normalization, and LeakyReLU nonlinearities. Each dataset was trained as fold 0 for 1,000 epochs using the standard nnU-Net v2 training recipe on a single H200 GPU. Training took approximately 20 seconds per epoch, or about 5.5 hours per dataset. On the held-out M&Ms-2 short-axis test set, the selected model achieved Dice scores of 0.944 for the LV cavity, 0.875 for LV myocardium, and 0.921 for the RV cavity.

At inference time, end-diastole and end-systole are first estimated efficiently rather than by segmenting the full stack at every cardiac phase. The estimator segments only the middle short-axis slice

across all phases and uses the LV cavity area curve as a phase-selection proxy: the phase with the largest LV cavity mask is selected as end diastole, and the phase with the smallest LV cavity mask is selected as end systole. Full-stack segmentation is then run only for these two selected phases. This two-stage procedure reduces computation while preserving the clinically relevant measurement frames.

The Quantitative Measurement Estimator converts the predicted full-stack masks into standard CMR-derived volumes and functional indices. Let $\mathcal{I}^{(p,z)}$ denote slice z at cardiac phase p, and let

$$g_\phi(\mathcal{I}^{(p,z)}) \in \{0, \mathrm{LV}, \mathrm{MYO}, \mathrm{RV}\}^{H\times W} \tag{13}$$

be the predicted mask for left-ventricular cavity, left-ventricular myocardium, and right-ventricular cavity. For class $c \in \{\mathrm{LV}, \mathrm{MYO}, \mathrm{RV}\}$, the physical area of class c on slice z and phase p is

$$A_{c,z}^{(p)} = \Delta x_z \Delta y_z \left| g_\phi(\mathcal{I}^{(p,z)}) = c \right|, \tag{14}$$

where Δx_z and Δy_z are the in-plane pixel spacings in millimeters. The phase-specific class volume is obtained by summing slice areas multiplied by the corresponding slice spacing,

$$V_c^{(p)} = 10^{-3} \sum_z A_{c,z}^{(p)} \Delta z_z, \tag{15}$$

where Δz_z denotes the through-plane spacing assigned to slice z, and the factor 10^{-3} converts cubic millimeters to milliliters. Under a uniform-grid representation, Eq. 15 reduces to multiplying the class voxel count by a single voxel volume; the slice-wise form is used to make the metadata dependence explicit.

The reported end-diastolic and end-systolic volumes are $\mathrm{EDV} = V^{(p_{\mathrm{ED}})}$ and $\mathrm{ESV} = V^{(p_{\mathrm{ES}})}$. Stroke volume and ejection fraction are computed as

$$\mathrm{SV} = \mathrm{EDV} - \mathrm{ESV}, \qquad \mathrm{EF} = 100 \times \frac{\mathrm{SV}}{\mathrm{EDV}}. \tag{16}$$

These definitions are applied to the LV and RV cavity volumes. LV myocardial mass is computed from the end-diastolic myocardium volume using a myocardial density of $1.05\,\mathrm{g/mL}$. Volumes and LV mass, but not EF, are additionally indexed to body surface area using the Du Bois formula (Du Bois & Du Bois, 1916),

$$\mathrm{BSA} = 0.007184 \cdot h_{\mathrm{cm}}^{0.725} \cdot w_{\mathrm{kg}}^{0.425}, \qquad X_i = X/\mathrm{BSA}. \tag{17}$$

Height and weight are read from scan metadata; if either value is missing or invalid, indexed quantities are treated as unavailable rather than being imputed. The final raw and indexed measurements are compared with sex-specific reference ranges to support the multiple-choice measurement questions. As a sanity check of the measurement pipeline on a case with available reference values, relative errors were 2.8% for LVEDV, 14.2% for LVESV, 10.1% for RVEDV, and 4.7% for RVESV.

Both evidence pathways were trained and evaluated on a single H200 GPU.

C LLM Implementation

Cardiomni uses different models for the trainable routing policy and the frozen language roles. The trainable planner is initialized from Qwen2.5-7B-Instruct and optimized with GRPO. The fixed planner helper, Evidence Sufficiency Verifier, and final answer generator are all served as frozen Qwen3-VL-4B-Instruct instances through an API-compatible vLLM interface. The visual and quantitative evidence backbones are not LLMs and are described separately in Appendix B. During the fixed-schedule ablation, the same frozen Qwen3-VL-4B-Instruct backbone is used for the single answer-generation call, with no learned planner, verifier loop, or GRPO update.

Role assignment. The agent runtime uses four language-model roles: planner-main, planner-fixed, verifier, and executor/generator. During GRPO training, planner-main is the only trainable role and uses Qwen2.5-7B-Instruct. The planner-fixed, verifier, and executor/generator roles use Qwen3-VL-4B-Instruct and remain frozen. For evaluation of the trained agent, planner-main is replaced by

the merged GRPO checkpoint at global step 175, served as `Qwen2.5-7B-grpo-ckpt175`; the other three roles remain frozen Qwen3-VL-4B-Instruct. In validation and test-time evaluation, generation is deterministic with temperature 0.0. During rollout sampling for GRPO, the logged vLLM rollout configuration uses stochastic sampling with temperature 1.0, top-$p = 1$, top-$k = -1$, and four sampled trajectories per prompt.

GRPO optimization details. The GRPO run uses a rule-based CMR reward backend and no learned reward model. Training data are stored as parquet files with 329,195 training rows and 8,133 validation rows. The logged training configuration uses train batch size 2, rollout group size 4, PPO mini-batch size 1, PPO micro-batch size per GPU 1, and log-probability micro-batch size per GPU 1. The actor learning rate is 1×10^{-6} with weight decay 0.01, constant warmup style, gradient clipping at 1.0, one PPO epoch per update, and no entropy bonus. The KL loss is enabled with coefficient 0.001, while KL is not added directly to the scalar reward. The clipping range is asymmetric, with lower clip ratio 0.2 and upper clip ratio 0.3. Gradient checkpointing and remove-padding are enabled for the actor model, and LoRA is not used (rank 0), so the planner checkpoint is a full-model update rather than an adapter-only update.

Context, response, and trajectory limits. The maximum prompt length is 18,432 tokens and the maximum response length is 2,048 tokens. The agent loop is capped at five evidence-collection steps during training, matching the main-method description, with an overall agent timeout of 500 seconds. The rollout engine uses vLLM with tensor parallel size 1 and bfloat16 rollout dtype. The logged run used one GPU for the actor/rollout process after local multi-GPU communication proved unstable on the workstation; earlier two-GPU configuration files used the same model-role split and hyperparameter choices except for the larger nominal train batch size.

Evaluation setup. For the reported GRPO checkpoint evaluation, planner-main is served from the merged step-175 Qwen2.5-7B checkpoint, while the planner-fixed, verifier, and executor/generator roles are served from Qwen3-VL-4B-Instruct. The VQA evaluation enables only the Visual Attention Highlighter, whereas the MCQ evaluation enables the Quantitative Measurement Estimator. The evaluation script allows up to eight agent steps and uses a maximum wall-clock time of 420 seconds per case, with deterministic decoding. Final answers are parsed by deterministic answer extractors: Yes/No for VQA and A–D for MCQ.

D DeepCMRVQA Dataset Design

D.1 Image Preprocessing

DeepCMRVQA preserves two image representations, because the visual-classification and quantitative-measurement pathways use different evidence forms. For the Visual Attention Highlighter, each patient is represented by up to five sequence groups: 2CH cine, 3CH cine, 4CH cine, short-axis cine, and short-axis LGE/PSIR. Frames are stored as grayscale JPG images and expanded to three channels before being passed to the 3D backbone. For each available sequence, 16 frames are selected by uniform temporal sampling; if fewer than 16 frames are available, frames are repeated by the same evenly spaced indexing rule. During classifier training, images are resized to 256×256 and randomly cropped to 224×224, with random horizontal flipping and ImageNet normalization. During validation, testing, Grad-CAM generation, and agent inference, the same resize is followed by a center crop to 224×224 and the same normalization. No physical voxel-spacing resampling is performed for this visual pathway; it operates on 2D frame clips intended for recognition and attribution rather than metric measurement.

For the Quantitative Measurement Estimator, preprocessing is geometry preserving. Short-axis cine DICOM slices are grouped by slice location and cardiac phase, then assembled into a 4D SimpleITK volume with the original in-plane pixel spacing, inter-slice spacing, origin, and orientation recorded from the DICOM headers. End-diastole and end-systole are selected using a fast phase-selection pass based on the LV-cavity area curve, after which full-stack nnU-Net v2 inference is run only on the selected phases. nnU-Net v2 applies its standard dataset-specific planning and preprocessing internally, including intensity normalization, resampling/cropping/padding as required by the learned 2D configuration, and 320×320 patch extraction. The final cardiac volumes are not computed from

normalized image coordinates: predicted masks are converted back to physical measurements using the DICOM-derived pixel spacing and slice spacing, as detailed in Appendix B.2.

E Additional Results

E.1 Ablation Study

Fixed-schedule baseline. We evaluate whether Cardiomni's adaptive Planner–Executor–Verifier loop provides benefits beyond a predetermined one-shot execution policy. The fixed-schedule baseline retains the same frozen Qwen3-VL-4B-Instruct backbone and the same task-specific evidence pathways as Cardiomni, but removes adaptive planning, evidence-sufficiency verification, iterative evidence acquisition, and GRPO-based policy optimization. Instead, every question follows a fixed execution path consisting of exactly one evidence request followed by a single answer-generation step.

For binary VQA, the fixed schedule always follows the Visual Attention Highlighter once. The predicted abnormality probability together with the corresponding Grad-CAM evidence maps are provided to the frozen backbone, which directly generates a binary Yes/No prediction. For quantitative MCQ, the fixed schedule always follows the Quantitative Measurement Estimator once. The measured value and its reference ranges are then provided to the same frozen backbone, which selects one of the four candidate answers. This design isolates the effect of adaptive planning while keeping both the perception modules and the answer-generation backbone identical to Cardiomni.

Binary VQA. Table 3 shows that Cardiomni achieves an overall balanced accuracy of 73.0%, substantially outperforming the fixed schedule (62.8%) by 10.2 percentage points. Cardiomni also achieves higher balanced accuracy on 13 of the 21 evaluated abnormalities, whereas the fixed schedule performs better on eight. The largest gains are observed for cardiac tumor (+41.3), pericarditis (+43.7), prior cardiac surgery (+33.6), no abnormal findings (+35.3), and myocardial edema (+24.0). The fixed schedule performs better on a small number of findings, including left ventricular wall motion abnormality (+18.5), left ventricular systolic dysfunction (+6.6), right ventricular dilation (+4.7), right ventricular systolic dysfunction (+5.7), prosthetic heart valve (+5.7), and delayed enhancement (+2.7). These results indicate that a single predetermined visual analysis is insufficient for many abnormalities, while Cardiomni's adaptive verification and iterative evidence acquisition substantially improve diagnostic robustness.

Quantitative MCQ. The benefit of adaptive reasoning is even more pronounced for quantitative measurement questions (Table 4). Cardiomni achieves 49.1% overall balanced accuracy compared with 22.4% for the fixed schedule, representing a 26.7-point improvement. Cardiomni outperforms the fixed schedule on 13 of the 14 measurement metrics, while the fixed schedule is marginally better only on right ventricular ejection fraction (+0.3). Furthermore, ten of the fourteen fixed-schedule scores remain at or below the 25% random-guess balanced-accuracy level. These results suggest that obtaining a single measurement is insufficient; reliable quantitative reasoning additionally requires verifying measurement quality and iteratively refining the evidence before mapping numeric values to discrete answer choices.

Taken together, these results demonstrate that Cardiomni's adaptive Planner–Executor–Verifier framework consistently improves both binary diagnostic reasoning and quantitative measurement interpretation. The gains are particularly pronounced for quantitative MCQ, where iterative verification prevents error propagation from imperfect measurements, but they also extend to binary VQA despite both methods using the same perception modules and frozen backbone. Since this ablation simultaneously removes adaptive planning, evidence verification, iterative evidence acquisition, and policy optimization, it evaluates the contribution of the adaptive decision-making loop as an integrated system rather than attributing improvements to any single component.

F Limitation

The fixed-schedule ablation does not isolate individual agent components. The ablation in §E.1 simultaneously removes learned routing, evidence-sufficiency verification, iterative evidence

Label	Cardiomni	Fixed schedule	Δ
aortic regurgitation	**67.3**	53.2	-14.1
bicuspid aortic valve	**68.0**	50.0	-18.0
cardiac tumor	**91.3**	50.0	-41.3
delayed enhancement	59.3	**62.0**	+2.7
left ventricular dilation	79.3	**81.5**	+2.2
left ventricular systolic dysfunction	68.7	**75.3**	+6.6
left ventricular wall motion abnormality	56.7	**75.2**	+18.5
mitral regurgitation	**66.0**	62.0	-4.0
myocardial edema	**74.0**	50.0	-24.0
myocardial infarction	**83.3**	62.0	-21.3
myocardial scar	**68.0**	66.7	-1.3
no abnormal findings	**85.3**	50.0	-35.3
pericardial abnormality	**74.7**	71.9	-2.8
pericardial effusion	**77.3**	56.9	-20.4
pericarditis	**90.7**	47.0	-43.7
prior cardiac surgery	**83.3**	49.7	-33.6
prosthetic heart valve	69.3	**75.0**	+5.7
right ventricular dilation	76.7	**81.4**	+4.7
right ventricular systolic dysfunction	70.0	**75.7**	+5.7
right ventricular wall motion abnormality	58.7	**62.4**	+3.7
tricuspid regurgitation	**66.0**	61.4	-4.6
Overall	**73.0**	62.8	-10.2

Table 3: Fixed-schedule ablation on binary VQA. Higher scores are in **bold**; Δ is fixed schedule minus Cardiomni, and – denotes an undefined score.

Metric	Cardiomni	Fixed schedule	Δ
left ventricular ejection fraction	**53.1**	27.3	-25.8
left ventricular end-diastolic volume	**50.4**	26.7	-23.7
left ventricular end-diastolic volume index	**44.8**	25.0	-19.8
left ventricular end-systolic volume	**52.7**	33.3	-19.4
left ventricular end-systolic volume index	**50.9**	16.2	-34.7
left ventricular mass	**49.3**	4.2	-45.1
left ventricular stroke volume	**47.4**	23.7	-23.7
left ventricular stroke volume index	**50.1**	25.0	-25.1
right ventricular ejection fraction	33.0	**33.3**	+0.3
right ventricular end-diastolic volume	**51.0**	15.6	-35.4
right ventricular end-diastolic volume index	**50.3**	25.0	-25.3
right ventricular end-systolic volume	**53.4**	21.2	-32.2
right ventricular end-systolic volume index	**46.2**	11.9	-34.3
right ventricular stroke volume index	**54.5**	25.0	-29.5
Overall	**49.1**	22.4	-26.7

Table 4: Fixed-schedule ablation on MCQ. Higher scores are in **bold**; Δ is fixed schedule minus Cardiomni, and random performance is 25.0.

acquisition, and GRPO training. It therefore measures the contribution of the adaptive loop as a bundle, rather than the independent effect of any one component. The fixed schedule's higher scores on several binary VQA labels show that learned routing is not uniformly beneficial when the appropriate pathway is predetermined by the question type. Separate comparisons that vary the routing policy, stopping rule, verifier, and training objective one at a time are needed to identify which component drives the large improvement observed on quantitative MCQ.

Evidence collection can terminate without sufficient evidence. As noted in §3.2, the Evidence Sufficiency Verifier's STOP/CONTINUE judgment is not enforced to reach YES: at the step budget $T_{\max}$, the Summarizer is invoked on whatever ledger exists, and produces a forced answer with no abstention option. A case where every evidence request failed, returned an out-of-scope error, or left the Verifier repeatedly unsatisfied is therefore indistinguishable, at the output interface, from one the Verifier assessed as sufficient. This also means that a case with plainly insufficient evidence would not naturally invalidate $\mathrm{cite}(\hat{y}) \subseteq h_{N+1}$ in Eq. equation 6: the citation constraint only guarantees the answer references *some* ledger entry, not that the entry it references is sufficient to support the claim made. Adding an explicit abstention path, and auditing what fraction of trajectories reach $T_{\max}$ without Verifier sign-off, are natural next steps.

The reward does not score the routing trajectory. As discussed in §3.3, R in Eq. equation 7 depends only on final-answer correctness $\mathbb{I}[\mathrm{parse}(\hat{y}) = y^\star]$, not on the routing sequence $\{s_1, \ldots, s_N\}$ or citation validity. A correct answer reached by a well-justified routing sequence is rewarded identically to one reached after a wasted or wrong evidence request that happened not to derail the final guess (e.g., a label with a strong prior toward one class, or a binary/four-way answer space with meaningful chance performance). This means the reported gains cannot currently be decomposed into "better routing decisions" versus "the underlying evidence pathways and answer space being informative enough that routing quality matters less than it appears"; §3.3 already flags this and proposes citation-validity and step-efficiency reward terms as future extensions, which we have not yet implemented.

The Visual Attention Highlighter's evidence is class-conditioned attribution, not independent detection. As clarified in §3.1, `attention_map` visualizes what drove an already-computed probability for a named label ℓ; it cannot propose or rule out a finding on its own, and every Grad-CAM overlay in the ledger presupposes a prior `predict`/`top_labels` call. Readers should not interpret the reported classification performance as evidence that the attention maps themselves are diagnostically discriminative – that claim would require an evaluation of localization quality against expert-annotated regions of interest, which this paper does not report.

Observed routing behavior includes repeated, non-productive requests. In qualitative inspection of training rollouts, we observed the policy occasionally re-entering an evidence pathway with a label string that does not exactly match the pathway's expected vocabulary (e.g., a capitalization mismatch), receiving the same validation error more than once within a single trajectory before recovering or exhausting $T_{\max}$. This does not appear as a distinct term in R (Eq. equation 7), since a trajectory can still end in a correct final answer despite such wasted steps; the step-efficiency reward proposed in §3.3 is intended to address this directly, but is not yet implemented or evaluated.